Lydia Fletcher

INF 385T

Astronomical Data: Long Term Preservation for Re-analysis

INTRODUCTION

Astronomy as a research discipline is a relatively small community that has tended to embrace standardization and open access initiatives at early stages in the development of those tools. Probably the most notable way in which astronomy adopted standardization was in the development and embracing of the FITS data standard. Historically, astronomers have tended to be careful with their data and, indeed, the OAIS Reference Model originally emerged from the efforts of the space science community.[1] However, like many other research disciplines, astronomy is facing a "data tsunami"[2] as data collections begin to creep from the terabyte range to the petabyte range. The Hubble Legacy Archive currently stores around 100 terabytes of data, but the upcoming Large Synoptic Survey Telescope will begin to generate data at "Large Hadron Collider levels" in the range of multiple petabytes per month as it surveys the entire sky every five nights.[3] Data sets that large will be far outside the human capacity for analysis and will require new techniques of automated extractions and analysis by the world's most powerful computers. In addition, the creation and funding of massive tera- and petabyte scale projects has led to a gradual reduction in the funding of smaller astronomical projects and the emergence of more international collaborative efforts. The effect that this trend will have on astronomy remains to be fully explored, but it may mean that astronomers conducting smaller researcher

---

[1] Gray, Norman and Graham Woan. "Digital Preservation and Astronomy: Lesson for funders and the funded." *ADASS XX* (2011): 1.
[2] Berriman, G. Bruce and Steven L. Groom. "How Will Astronomy Archives Survive the Data Tsunami?" *Communications of the ACM* 54:12 (December 2011): 1.
[3] Gaffney, Niall. Personal interview. 14 February 2013.

projects will move towards less original data collection and more re-use of archived data from the data generated at large missions.[4]

Alongside the multiple challenges such enormous data sets pose in terms of data curation and analysis, astronomy faces another massive data problem as it enters an era where the re-use of archival data becomes as important as the analysis of newly-captured astronomical data. The potential rampant loss of archival data is sparking new debates about data preservation in astronomy alongside the discussions about how to manage new massive data collections. In this essay, I will explore the current data trends and practices in astronomy and examine the emerging movements in data preservation and re-analysis, particularly in the context of data formats, practices, and repositories.

ASTRONOMICAL DATA: FORMATS AND STANDARDS

The most widely-accepted data standard for astronomical data is the Flexible Image Transport System (FITS). FITS was first described by D. C. Wells, E. W. Greisen, and R. H. Harten in 1981 in response to the massive increase in digital images produced by new astronomical observational equipment. The astronomers who created the FITS standard determined that astronomical discoveries were being hampered by the lack of a standardized system of sharing image files in the astronomical community: they found that "almost every installation has developed at least one unique data format and produced a large quantity of software based on the use of that internal format."[5] A single standard was needed to enable the sharing of astronomical data between observatories and researchers. Initially, the National Radio

---

[4] Wynholds, Laura, David Fearon, Christine L. Borgman, and Sharon Traweek. "When use cases are not useful: Data practices, astronomy, and digital libraries." *Publications, UCLA Graduate School of Education and Information Studies, UC Los Angeles* (2011): 1.
[5] Wells, D. C., E. W. Greisen, and R. H. Harten. "FITS: A Flexible Image Transport System." *Astronomy and Astrophysics Supplement Series* 44 (1981): 363.

Astronomy Observatory, the Netherlands Foundation for Radio Astronomy, and the Kitt Peak National Observatory adopted FITS,[6] but it has come to be the predominant system for storing and transferring astronomical data. Originally, FITS was developed to store information on seven-track and nine-track magnetic tape,[7] but over time it has evolved to be a digital format accessible by a wide variety of operating systems and programs created in a wide variety of programming languages.[8] Although FITS is an international standard, most astronomers still develop their own software to analyze FITS based on their research projects and uses of FITS files.

FITS was created to be a standard file format that contained all the data relevant to astronomical research in a single file. It was designed to be a single, flexible, powerful format to transmit the most common astronomical data of the time, the *n*-dimensional, regularly spaced data array,[9] providing for the transfer of images with accuracies up to 32 bits.[10] FITS was also designed to allow for unlimited and unrestricted addition of ASCII textual information within a human-readable FITS header—this enabled the transfer of information such as searchable keywords and what we would consider metadata about the creation, structure, coordinate systems, and any auxiliary parameters of the *n*-dimensional, regularly spaced data array.[11] As such, FITS is not a format for viewing rasterized images and, indeed, many FITS files do not contain image data at all. FITS is used primarily for transferring images that are ready for

---

[6] Wells et al 364.
[7] Wells et al 364.
[8] The FITS Support Office *FITS I/O Libraries*, http://fits.gsfc.nasa.gov/fits_libraries.html
[9] Wells et al 367.
[10] Wells et al 363.
[11] Wells et al 364.

analysis using specialized software but not meant to be viewed by human eyes: images must be processed and converted to a rasterized file before they can be viewed by humans.[12]

The FITS format was designed to be backwards compatible, so that as new features or standards are developed they can be deployed backwards to older FITS files as well as forwards to new ones.[13] Over time, the integer support within FITS has been increased from the original 32-bits to 64-bits as of the latest release of FITS standards, Version 3.0.[14] In addition, an elaborate set of FITS conventions called the World Coordinate System was defined to specify the physical, or world, coordinates to be attached to each pixel of an *n*-dimensional image.[15] The World Coordinate System was implemented to enable researchers to map an element in a data array to standard physical coordinates on the sky.[16] FITS continues to be an organic file format that is constantly evolving to suit the needs of astronomical researchers.

FITS was designed with long-term storage and accessibility in mind, which is expressed most succinctly in the maxim "once FITS, forever FITS." In addition, the developers and users of FITS have recognized the need for ensuring data quality and provenance as an essential part of the collection, use, and re-use of astronomical data by astronomers and astrophysicists, and most researchers make extensive use of the keyword/value header space within FITS files to fully describe the contained data. Because FITS relies on an extensive set of reserved keywords and controlled value lists, as well as the ability to add project-specific keywords and values, FITS files often have high quality associated metadata contained within the FITS file. In addition, the community of researchers who use FITS have developed a practice of making use of the history

---

[12] NDIIPP *Sustainability of Digital Formats Planning for Library of Congress Collections: Flexible Image Transport System (FITS), Version 3.0*, http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
[13] http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
[14] http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
[15] The FITS Support Office *FITS World Coordinate System (WCS)*, http://fits.gsfc.nasa.gov/fits_wcs.html
[16] http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml

and comment records within FITS files to store even more data description.[17] One current trend in the preservation of astronomical data is the conversion of historical data to FITS files so that they can be stored and used for research alongside more recently produced data. In March 2010, the Vatican Library announced that digitized scans of historical books and manuscripts on astronomy within the Vatican's collections would be converted and stored in the FITS format.[18]

The use of FITS is governed by the FITS standard, which was published and maintained by the NASA/Science Office of Standards and Technology, part of the National Space Science Data Center run by NASA.[19] The FITS standard outlines the requirements for the development of data to be stored in a FITS format, and specifies that an "archival format must be utterly portable and self-describing, on the assumption that, apart from the transcription device, neither the software nor the hardware that wrote the data will be available when the data are read."[20] In addition to describing in detail the way in which FITS files should be formatted, the standard is also responsible for creating and maintaining the controlled vocabulary of reserved keywords for use across the discipline: there are fifty-three keywords currently defined in the standard,[21] including details about which keywords are required and which are optional in the various sections of the FITS file.[22] In addition to the reserved keywords identified by the FITS standard, other agencies maintain databases of project-specific keywords and standard descriptions, including the Space Telescope Science Institute and the European Southern Observatory.[23]

---

[17] http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
[18] http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
[19] Space Telescope Science Institute *Definition of Flexible Image Transport System*, http://archive.stsci.edu/fits/fits_standard/
[20] Space Telescope Science Institute *Definition of Flexible Image Transport System: Overview*, http://archive.stsci.edu/fits/fits_standard/node5.html#SECTION00500000000000000000
[21] The FITS Support Office *FITS Standard Dictionary*, http://heasarc.gsfc.nasa.gov/docs/fcg/standard_dict.html
[22] http://archive.stsci.edu/fits/fits_standard/
[23] The FITS Support Office *FITS Keyword Dictionaries*, http://fits.gsfc.nasa.gov/fits_dictionary.html

At the start of the twenty-first century, the astronomical community began to recognize the importance of using archival data as a tool for conducting new research. When planning the for the long-term life cycle of the Hubble Legacy Archive, the Second Decade Committee at the Space Telescope Science Institute found that the rate at which scientists were withdrawing data from the archive for re-analysis was four times greater than the rate at which data was being ingested into the archive[24] and by 2005 the Space Telescope Science Institute recognized that overwhelmingly more papers were being published that drew on archival data than newly collected data.[25] The increasing importance of the use of archival data in research led to the creation of the Virtual Observatory and the International Virtual Observatory Alliance, a powerful collection of international interoperating data archives, software tools, and metadata standards that use the Internet to create a vast system of interlinked environments for sharing and analyzing astronomical data.[26] Most astronomical data of value is the highly processed data that typically resides only on departmental servers and the personal computers of the astronomers using the data for research. Most often, the data that generates new discoveries in astronomy is not published alongside the graphical representations provided in journal articles. The awareness of this lack of access to archival data and the threat of data loss when astronomers move between institutions or retire is driving the desire for a unified approach to archiving and preserving data for long-term re-analysis.[27]

In the United States, the Virtual Astronomical Observatory is the gate through which US astronomers engage with other International Virtual Observatory Alliance members around the

---

[24] The Hubble Second Decade Committee. *The Hubble Data Archive: Towards the Ultimate Union Archive of Astronomy* (Baltimore, MD: 2000), 4.
[25] Gaffney, Niall. Personal interview. 14 February 2013.
[26] International Virtual Observatory Alliance *What is the VO?* http://www.ivoa.net/about/what-is-vo.html
[27] Norris, Ray, Heinze Andernach, Guenther Eichhorn, Françoise Genove, Elizabeth Griffin, Robert Hanisch, Ajit Kembhavi, Robert Kennicutt, and Anita Richards. "Astronomical Data Management." *Highlights of Astronomy* 114 (August 2006): 7.

world. Supported by the High Energy Astrophysics Science Archive Center, the National Optical Astronomy Observatory, the National Radio Astronomy Observatory, the Smithsonian Astrophysical Observatory, and the Space Telescope Science Institute, among others, the US Virtual Observatory works with US-based astronomical research missions to develop and provide new and improved research tools and capabilities to the US astronomical community.[28]

In more ways than one, astronomy represents an ideal discipline to begin an international exploration of the curation and archiving for long-term access of scientific research data. The astronomical community is small, most data is already contained in a discipline-wide standard format, and the members of the community are highly aware of ePublishing and supportive of using initiatives such as the Astrophysics Data System to disseminate new knowledge.[29] The goal of the Virtual Observatory is to provide astronomers around the world with seamless access to online resources by defining interoperability standards such as those for resource registries, query languages, data access layers, content descriptions, and data models.[30] Many of the participants in the International Virtual Observatory Alliance are knowledgeable about data archives and strive to provide "useful and usable" standards so that data and software can be "published" in the Virtual Observatory and accessible by other astronomers.[31] The Virtual Observatory describes itself as "a kind of club of data services that all follow the same rules."[32]

One of the ways in which the Virtual Observatory are contributing to the seamless sharing of astronomical data are the creation of new standard data formats that builds on the

---

[28] The US Virtual Astronomical Observatory *About the VAO*, http://www.usvao.org/about-vao/
[29] Norris et al 7.
[30] Norris et al 2.
[31] Norris et al 2-3.
[32] The International Virtual Observatory Alliance *Using the VO*, http://www.ivoa.net/astronomers/using_the_vo.html

traditional FITS files,[33] and the creation of new standards such as the Astronomy Data Query Language.[34] The VOTable format is an XML standard that was created to improve the interchange of data held in sets of tables. Although FITS can be used to store and transfer data held in tables, VOTable improves upon the FITS Table format by mirroring the FITS Binary Table format in XML. The primary way in which VOTable improves on FITS is its adoption of the streaming file transfer method—because of the way FITS files are built and accessed, with much of the metadata held in a rigid header structure, FITS is a difficult file format to stream over the web. Because no FITS library is required to interpret the data stored in a VOTable format, VOTables work well in the new streaming data paradigm created by the rise of the Internet in sharing scientific data.[35] The streaming data model is particularly robust for large-scale observatory projects that make use of distributed computing models to analyze collected data.[36] The VOTable format standards also include powerful new metadata hierarchies, which standardize the parameters of data stored in VOTables so that they are easily sharable and translatable internationally.[37] The VOTable format was designed so that FITS files can be converted to VOTables and vice versa without loss of data.[38]

The establishment of the Virtual Observatory has also led to an examination of the traditional concept of the "data center" within the astronomical community. The Virtual Observatory has led to the definition of a "data center" in terms of the services and value a data center can provide to its community, as well as the added-value of data management expertise,

---

[33] http://www.ivoa.net/astronomers/using_the_vo.html
[34] The International Virtual Observatory Alliance *VO Glossary for Astronomers*, http://www.ivoa.net/astronomers/vo_glossary.html
[35] Ochsenbein, François and Roy Williams. "VOTable Format Definition." *IVOA Proposed Recommendation 2013-03-15 (Volute Revision: 2028)*: 5. http://www.ivoa.net/Documents/VOTable/20130315/
[36] Ochsenbein et al 8.
[37] Ochsenbein et al 6.
[38] Ochsenbein et al 8.

sustainability, and quality of the data stored in the center. The Virtual Observatory's distributed model incorporates information and services stored in a variety of "traditional" data centers such as observatory archives, discipline-specific data centers, and separate institutions which provide reference services and tools. The Virtual Observatory, therefore, is shifting emphasis towards observation archives (with an emphasis on providing "science ready data"), value-added services and tools, theoretical services such as on-demand analytical services, software suites, and fully realized virtual research environments.[39]

The nature of the Virtual Observatory is also leading to a re-evaluation of the nature of data curation in the astronomical community, while recognizing that quality data curation comes with increasing costs for selecting, homogenizing, describing, and distributing data so that it is "science ready" through the Virtual Observatory. However, even though data providers will have to increase the amount of curation resources they commit to their data, doing so will result in higher quality data and attendant metadata which will, in turn, result in better services provided to the community at large and, theoretically, more occasions to collaborate with colleagues, thereby increasing their visibility in the community and the usage of their data.

One US Virtual Astronomical Observatory initiative that began in 2007 seeks to tackle the gap in data preservation in astronomy recognized that although many observatories and departments are beginning to create archives for their data or store their data in digital repositories, often the most highly processed data on which journal articles, conference papers, and other major published analyses are based is not archived or stored in a permanent

---

[39] Norris et al 3.

repository.[40] Astronomers have increasingly begun to demand better links between their publications and the data behind them, where the term "data" is taken to include primary observational data, published results based on those data, and graphical representations of results.[41] The team behind the US VAO initiative aims to use Fedora, a flexible and powerful open source repository software, to create a set of web services that interlink published literature with the associated astronomical data sets. Fedora was designed to support the association of "behaviors" with digital objects, so that datasets and FITS files can be ingested into the database and treated as single objects with journal articles and conference papers associated with them as "behaviors."[42] The data and articles will also be interlinked with journal and Virtual Observatory portals, which will greatly increase data discovery and accessibility by allowing researchers to search for, for example, "images of NGC 4151" and return a list of results.[43] The project envisions using a Dublin Core metadata framework to meet the challenge of creating high quality metadata to support discovery, with modifications to simple Dublin Core tailored to the needs of the astronomy community, which will be made available through the Open Archives Initiative Protocol for Metadata Harvesting.[44]

Much work remains to be done, and many challenges remain, especially with regard to funding, long-term access, high quality metadata standards, data curator certification, version control, and intellectual property standards.[45] The astronomical community is striving to create "best practice" standards through the implementation of the Virtual Observatory and is moving

---

[40] Choudhury, Sayeed, Tim DiLauro, Alex Szalay, Ethan Vishniac, Robert Hanisch, Julie Steffen, Robert Milkey, Teresa Ehling, and Ray Plante. "Digital Data Preservation for Scholarly Publications in Astronomy." *The International Journal of Digital Curation* 2:2 (2007): 24.
[41] Norris et al 5.
[42] Choudhury 26.
[43] Choudhury 27.
[44] Choudhury 28.
[45] Norris et al 4.

forward in unison, but the discipline still lacks a strategic data framework that is actively developing data management policies or a strategy for representing the astronomical community and its interests to external parties.[46] There has been a push to develop robust education and public outreach programs to help to convey the results of astronomical research to the general public, but there is still a great deal of work to be done.

Preservation and Re-analysis of Astronomical Data

As discussed previously, many institutions that archive and preserve astronomical data are discovering that an increasingly large percentage of new astronomical discoveries and being made using archival data that was collected and stored by missions that have now ended. Arguably the most famous telescope mission in the world is the Hubble Space Telescope, which was launched by NASA in 1990 and which has proven to be "one of NASA's most successful and long-lasting science missions."[47] The Hubble has collected hundreds of thousands of images and a collection of data that has changed both the fundamental understandings that astronomers have about our universe in addition to the general public's understanding of what astronomical missions comprise. As the Hubble mission entered its second decade, a committee was formed to look to the future of Hubble's data and the Hubble Legacy Archive was created to ensure the long-term preservation and accessibility of the data captured by Hubble. When the Hubble Second Decade Committee released their initial report in December, 2000, they already recognized that researchers were accessing and using more archival data than was begin ingested into the archive from new Hubble captures.[48] In addition, the open access policy in effect at the Hubble Legacy Archive has resulted in approximately three times as many papers (and citations)

---

[46] Norris et al 10.
[47] HubbleSite *Hubble Essentials*, http://hubblesite.org/the_telescope/hubble_essentials/
[48] The Hubble Second Decade Committee 4.

created using archival data retrieved from the Hubble archive as those based on the original Hubble data.[49] Now that the Hubble mission is officially entering its final collection phase, the Hubble Legacy Archive is facing a new era in which it will be a static data archive that will see 100% of its data use fall into a new model of discovery based on archival data usage.[50]

Even as astronomers are working to develop international standards, services, and software through the International Virtual Observatory Alliance and to probe the issues surrounding the best ways to store data and make it accessible to researchers, the question remains of what the best mode for long-term data storage and access is. Researchers across the discipline are beginning to understand the importance of having access to as much data as possible on the objects or phenomena they are studying. Although astronomers have always enjoyed a tradition of open access, astronomical data centers, which typically provide access to data for all astronomers at no charge,[51] are beginning to embrace an archival model reminiscent of that typically employed in libraries and archives and to develop repositories that have more of an emphasis on preservation with a purview to providing access to historical data for re-use in new research projects. One such project was proposed by the Pisgah Astronomical Research Institute in Rosman, North Carolina: after a survey of historical photographic plates containing capture astronomical images found that the plates were receiving little usage, the Pisgah Institute volunteered to create an archive of the plates and to digitize them so that they would be discoverable and accessible to astronomical researchers. The plan to store and digitize the plates, which date from the 1880s up to the 1990s, has met with a fair amount of critics who believe the funding contributed to the moving and storing of the plates could be better spent on new research

[49] Norris et al 5.
[50] Gaffney, Niall. Personal interview. 14 February 2013.
[51] Norris et al 4.

projects. However, the a group at Harvard that has been digitizing their plates have already made several new discoveries and it is believed that having the historical data to compare with newly collected data will enable researchers to track changes in star sizes and galactic motions. The Pisgah Institute has an important outreach and education component, particularly to K-12 programs, and their aim is much in line with the missions of most research libraries and archives: to preserve data and make it accessible to users.[52]

Can scientific disciplines use data curation management techniques and tools developed for libraries and archives? A team at the University of California-Los Angeles, in conjunction with the Data Conservancy, began work on an 18 month survey on astronomy data practices with funding from the National Science Foundation's DataNet Initiative grant-funding organization. The research was conducted with especial emphasis on examining their methods for use, preservation, and sharing from an Information Studies standpoint. The UCLA team approached astronomical data with three questions:

1. What are the data management, curation, and sharing practices of this community?
2. Who uses what data when, with whom, and why?
3. What data are most important to curate, how, and for whom?[53]

By taking such an approach, the research team at UCLA explored what astronomers at major sky survey missions consider to be their data, their criteria for data selection, their criteria for data trust, and the various problems they have encountered in managing their data.[54] The goal in examining astronomical data from this perspective was to attempt to understand what form an astronomical digital library could take that would ensure long-term preservation and access. The

---

[52] Feder, Toni. "North Carolina institute offers to archive old astronomy data." *Physics Today* 62:3 (March 2009): 26-7.
[53] Wynholds et al 1.
[54] Wynholds et al 1.

initial report by the UCLA team indicates that there are a wide variety of issues around data production, use, preservation, and sharing that need to be addressed, and that many astronomers' concerns revolve around the initial source of the data they want to use and, as has been found by the Virtual Observatory, that many astronomers have concerns about the value-added services that are available alongside the data. Increasingly, a model for scientific data storage for preservation and access could be based on emerging trends in digital libraries, because digital libraries "hold the potential to move beyond merely disseminating resources toward creating environments that support the analysis required to understand them."[55]

One of the most important initial findings in the UCLA team's report is the conclusion that trust plays a major part in how astronomers choose archival data to re-use for new projects. Astronomers take the trust that they put in the mission that initially collected the data into account in making decisions, with well-established, large-scale projects such as the Sloan Digital Sky Survey or the Hubble Space Telescope mission (and attendant Hubble Legacy Archive) representing highly trusted sources because the data they collect are extensively vetted and publicly reviewed, and the instruments with which the data are collected are well calibrated. They represent sources that are "valid, accurate, well documented, and trustworthy."[56] Astronomers who were interviewed during the research project referred to data sets that had been curated from trusted sources such as a NASA archive or the Sloan Digital Sky Survey as "secondary data products," and some astronomers expressed the sentiment that they knew who to trust in the field as a producer of good data that could be re-used without worrying about the trustworthiness of the data.[57]

---

[55] Wynholds et al 1.
[56] Wynholds et al 2.
[57] Wynholds et al 2.

The implications that this has for developing digital library-like data archives for astronomical data include the need to involve considerable human expertise to ensure that data curation and the development of data sources that astronomers can trust. Information Studies professionals have been working to develop and expand the concept of the trusted digital repository for over a decade. Trusted digital repositories should provide reliable, long-term access to curated and managed digital sources for the foreseeable future.[58] The aim in creating a trusted digital repository is to ensure that users understand that the information housed in the repository is of high quality, expertly curated, reliable, and trustworthy. All of these are attributes that the research team at UCLA found to be of importance to astronomical researchers when evaluating data sources. The data archives currently viewed most favourable by astronomers, with guidelines created by the International Virtual Observatory Alliance in conjunction with Information Studies professionals with expertise in creating trusted digital repositories could go a long way to enabling astronomical researchers to acquire and re-use high quality, "science ready" data for new, small-scale projects.

CONCLUSION

Like many other scientific research disciplines, major new advances in the tools and techniques that astronomers are using to capture their research data are beginning to produce massive data sets that are generating new challenges with regards to collection, storing, curating, archiving, preserving, and re-using the data. Unlike most other disciplines, however, astronomy is well-equipped and well-placed to handle these challenges because of the pre-existing values and standards in place within the discipline. The almost-universal usage of the FITS data format and the standards associated with it enable astronomers to archive and share their data with each

[58] OCLC-RLG. *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, CA: 2002) i.

other easily. In addition, the prevailing positive attitudes within the discipline towards open access initiatives mean that most members of the community favor describing and storing their data in a manner that makes it easy for other researchers and the general public to access it. The International Virtual Observatory Alliance was formed and designed to tackle just such data management issues in a manner that will ensure uniformity around the world and by looking towards established concepts of what a digital repository should be from the world of libraries and archives, astronomers can begin to develop trusted digital union archives that store and preserve data for long-term access and re-use in ways that meet the expectations of astronomers for trustworthy data.

[Word count: 4,697]